A Study on Life Aspect Inference based on Association with Latent Topics

氏　名	山本 修平
学 位 の 種 類	博　士（ 情報学 ）
学 位 記 番 号	博 甲 第　7891　号
学位授与年月日	平成　28 年　3 月　25 日
学位授与の要件	学位規則第 4 条第 1 項該当
審 査 研 究 科	図書館情報メディア研究科
学位論文題目	A Study on Life Aspect Inference based on Association with Latent Topics

主	査	筑波大学　教授	博士（工学）	佐藤哲司
副	査	筑波大学　教授	理学博士	北川博之
副	査	筑波大学　教授	博士（工学）	歳森 敦
副	査	筑波大学　教授	博士（工学）	森嶋厚行
副	査	筑波大学　准教授	博士（情報学）	手塚太郎

論　文　の　要　旨（2、000 字程度）

現在、知識共有コミュニティサイトやブログ、マイクロブログなど、多くの情報共有サービスが存在している。ツイッターは最も広く普及しているマイクロブログの 1 つであり、ユーザの経験や意見、また日常生活でのイベントなど、身近でかつ新鮮な情報がツイートとして投稿されている。本研究では、地域性が高く新鮮、かつ他ユーザに有益なツイートを「実生活ツイート」と呼んでいる。実生活ツイートは生活の様々な局面に対応している。たとえば、「電車が来ない」というツイートは生活の中の「交通」の局面に対応し、これから電車に乗ろうとしているユーザを支援できるであろう。「雨が降ってきた」というツイートは「気象」の局面に対応し、これから外出する人や、洗濯しようとする人など、幅広いユーザを支援できるといえる。本研究では、様々なポータルサイトや Wikipedia のカテゴリ構成などを参考に、人々の生活を典型的な 14 の局面に整理し、実生活ツイートが関連している局面を推定する手法を論じている。

未知のツイートに対する局面推定における技術的な課題は、実生活ツイート以外に「ありがとう」や「そうなんだ」などの相槌や共感なども多く投稿されていること、ツイート本文が高々数十文字と文字数が少ないこと、ツイートに関連する局面は一つとは限らず、可変個数の局面を推定しなければならないことである。特に、推定する局面の数は、最も関連が強い一局面だけを推定して欲しい場合から、出来るだけ幅広く少しでも関連すると思われる局面は網羅的に推定して欲しい場合まで、ユーザの利用形態に依存して大きく変動すると考えられる。このため、局面推定に利用できる手がかりが少ない状況においても、可変個数の局面を推定できるダイナミックな推定手法が必要とされていた。

本研究では、上述の課題を解決するために、潜在的なトピックと局面の対応関係に基づく階層的推定法を提案している。階層的推定法の基本的なアイデアは、教師なし学習と教師あり学習を階層的に組み合わせ、両者の効用を融合する新たな学習手法を実現することにある。第 1 段階では、教師なし学習として知ら

れる潜在的ディリクレ配分法（LDA）を用いて、大量のツイート集合からトピックを抽出する。このトピックは、ツイートが持つ潜在的なトピックであると考えることができる。第 2 段階では、局面ラベルが付与された少量のツイートを用いて、抽出した潜在トピック（以下、トピックと称する）と局面の関連度を算出し、局面に複数トピックを結びつけた対応関係を構築する。実際に未知のツイートに局面を推定する際は、ツイートに出現する単語から、その単語の出現するトピックの生起確率とそのトピックが対応関係を持つ局面への関連度を用いて、局面毎にスコアを算出する。

これまでも、Naive Bayes 分類器や SVM(Support Vector Machine)など教師あり機械学習によるラベル推定手法が知られており、両手法とも複数のラベルを推定するマルチラベリングへと拡張されている。またトピックモデルの 1 つである Labeled LDA もマルチラベリングを目的に提案されている。いずれの手法も十分な訓練データを用いることで、ブログやニュース、Web ページなどの比較的長い文書の分類を高い推定精度で実現している。しかし、本論文で課題とする、短文かつ訓練データが少ない場合には、考慮できる手がかり語が少なく十分な性能が得られていない。この原因として考えられることは、従来の教師あり機械学習は、訓練データから直接クラスラベルに対する単語の尤度を学習していることである。提案する階層的推定法は、ツイートに出現する単語を潜在的なトピックに展開し、ツイートが言及している話題をトピックという単位で確率的に拡張した後に、少量の訓練データでトピックと局面の関連度を算出しツイートに局面を推定することを特徴としている。この確率的な拡張に LDA を用いることで、推定精度の低下要因であるノイズとなる単語の混入を避ける工夫をしている。

本論文で述べているマルチラベル分類においては、スコアが閾値を超えた局面をツイートに付与することにより実現している。第 4 章では、多くの局面が同じトピックに対して結びつく競合問題を解決するために、Entropy Feedback を階層的推定法の第 2 段階に導入している。Entropy Feedback は、ある時点のトピックと局面の対応関係に対してエントロピーを算出し、その値からフィードバック係数を求めて関連度を再算出する繰り返し演算からなる。この Feedback 機構によってトピックと局面の対応関係を洗練し、推定性能の向上を目指している。京都市内で投稿された日本語ツイートを用いた評価実験を行い、階層的推定法は未知のツイートに適切に局面を付与できることを明らかにした。Entropy Feedback を導入した提案手法は、それぞれの局面に特徴的なトピックが強い関連度で結びついており、対応関係が洗練されるプロセスを確認している。従来のマルチラベル分類手法と適合率、再現率、F 値を用いて推定性能を比較した結果、階層的推定法は高い F 値を示していた。特に、訓練データの数を減らした場合には、従来手法は推定精度が低下していたが、提案手法はほとんど下降しないという特徴も明らかにしている。

第 5 章では、ユーザの指向に合わせた柔軟な個数の局面推定を、未知のツイートに対して生起する局面を確率分布として推定することで実現している。ここでは、ラベルが付与された訓練データでモデルを学習し、入力された未知のツイートに対して、t 検定を用いて最適なトピックと局面の対応関係を確率分布として推定する手法を提案している。評価実験の結果、提案した t 検定に基づいた局面の確率的な推定手法が、ベースライン手法に比べて高い推定性能を示していた。訓練データに単一ラベルを付与した場合と、複数ラベルを付与した場合で、JS 情報量（Jensen-Shannon Divergence）によって確率分布の推定性能を評価した結果、特に単一ラベルという状況で階層的推定法は有意に良い推定ができることも明らかになった。

以上の結果から、ツイートのような短文に対して、より少ない訓練データでマルチラベル分類をする場合や、確率分布推定をする場合に、提案した階層的推定法が有効であることが明らかになったといえる。

審　査　の　要　旨（2、000 字以上）

【批評】

ツイッターなどのソーシャルメディアの普及発展は著しく、生活者が投稿した日常に関する情報から、生活に有益な知識を抽出する技術の発展深化に資する研究は喫緊の課題となっている。本研究は、人々の実生活を支援できると考えられる実生活ツイートと相槌などの非実生活ツイートが混在して投稿されているツイッターを対象に、実生活ツイートを抽出し関連する実生活の局面をラベル付けすることを課題としており、時宜を得た研究といえる。

本研究の基本的なアイデアは、あらかじめ用意した正解データで学習を行ってモデルを構築し、モデルに基づいて未知のツイートにラベルを付与する機械学習の範疇に入る技術である。ここで、平均文字数が高々数十文字と短く学習の手がかりが少ないツイートを対象としていること、更に、人手での作成が避けられない正解データが少量であっても高い性能を発揮することを目指しているところに特徴がある。論文の構成は、第 2 章で関連研究を概観した後に、第 3 章で提案する階層化推定法の枠組みを詳細に述べている。第 4 章では、未知のツイートに複数の局面をラベルとして付与する手法として、階層的推定法の第 2 段階に Entropy Feedback を導入する手法を提案し、実際のツイートデータを用いてその有効性を論じている。更に、第 5 章では、局面の推定をラベルの有無ではなく、確率的なラベル付与問題として論じている。第 6 章では、これまでの議論を踏まえて、本論文の到達点を明らかにしており、最後に第 7 章でまとめを行っている。以下、各章毎に議論を概観し批評を行う。

第 1 章は、マイクロブログの一実現であるツイッターが普及している現状を明らかにするとともに、東日本大震災での経験などの実例を踏まえて実生活を支援できることを具体的に述べている。本論文が目標としているツイートに生活の局面を推定するという課題の妥当性は十分に説明されており、研究の意義は十分に認められる。

第 2 章は、上記課題に関連する先行研究を、ツイッターからの情報抽出、経験マイニングなどの応用面からと、トピックモデルやマルチラベル分類などの技術面から概観・整理し、本研究の位置づけを明らかにしている。本論文をまとめる範囲で十分な調査をしていると言えるが、近年の機械学習に関わる技術領域の急激な発展を鑑みると、技術面からの調査は継続的に行っていくことが不可欠と言える。

第 3 章は、本論文の核となる階層的推定法のフレームワークを論じている。大量のツイートを入力として教師無し機械学習として知られている LDA を適用して潜在的なトピックを抽出する第 1 階層と、あらかじめ指定された実生活の局面（論文では 14 局面）との対応関係を構築する第 2 階層とで構成するとしている。本論文の後半（第 4 章、第 5 章）は、第 2 階層の実現方法に関する議論で有り、第 1 階層に関する議論は、この章に限定されている。実績のある LDA の適用は概ね妥当であると言えるが、提案する階層化推定法における第 1 階層は教師無し機械学習であることだけが条件と思われる。LDA 以外にも多くの教師無し機械学習は知られていることから、これらの手法との組み合わせの可能性についても論じられていると、提案法の有効性が更に明確に示せたと思われる。また、2 階層をもって階層的と称しているが、3 階層以上への拡張の可能性についても論じることができれば、更に本質的な議論へと踏み込むことができたと思われる。

第 4 章は、第 3 章で述べた階層的推定法の第 2 階層に Entropy feedback の機構を導入することで、個々のツイートに対して適切な数の局面を動的にラベル付けするマルチラベル分類手法を提案している。物理学の分野などでよく知られている Entropy という概念を情報学の分野に導入し、期待する効果を得ようとする試みは新規性が有り、高く評価することができる。マルチラベル分類における Entropy 算出式を導出し、feedback 機構による局面とトピックの最適な関連づけを実現し、実データを用いた評価によって有効性を検証する論旨の展開に問題は感じられない。

第 5 章は、ツイートに関連する局面を確率的に推定しようとする提案である。これは、明日の天気を「晴れ時々曇り」としていた天気予報を「降水確率〇%」とすることに相当する。推定結果の提示法としては大きな変革となっているが、第 4 章に示したマルチラベル推定の自然な拡張として理解することができる。ナイーブベイズ分類や SVM(Support Vector machine)、Labeled-LDA など既存の分類手法と比較をする過程で、局面によって推定のされ方に違いがあることや、人手による推定では見逃している可能性のある局面の存在を示唆するなど、提案法に関して多くの知見を得ることに成功している。

第 6 章の考察、および第 7 章のまとめでは、第 1 章で示した本論文の課題に対して、本研究の取り組みが明らかにした範囲を論じている。本研究の遂行によって設定した課題が解決できていることを示すとともに、新たに発掘された課題の存在も示唆されている。

以上を総合的に判断すると、本論文は情報学の学位論文として十分な内容を含んでいると認められる。

【最終試験結果】

平成 28 年 1 月 25 日、図書館情報メディア研究科学位論文審査委員会において、審査委員全員出席のもと、本論文について著者に説明を求めた後、関連事項について質疑応答を行った。引き続き、「図書館情報メディア研究科博士後期課程（課程博士）の学位論文審査に関する内規」第 23 項第 3 号に基づく最終試験を行い、審議の結果、審査委員全員一致で合格と判定された。

【結論】

よって、本学位論文の著者は博士(情報学)の学位を受けるに十分な資格を有するものと認められる。